

Randomized Structural Sparsity via Constrained Block Subsampling for Improved Sensitivity of Discriminative Voxel Identification

Yilun Wang^{a,b,c}, Junjie Zheng^b, Sheng Zhang^a, Xunjuan Duan^b, Huaifu Chen^{b,*}

^a*School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731 P. R. China.*

^b*Key laboratory for Neuroinformation of Ministry of Education, School of Life Science and Technology, and Center for Information in Biomedicine, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611054, P. R. China.*

^c*Center for Applied Mathematics, Cornell University, Ithaca, NY, 14853, USA*

Abstract

In this paper, we consider voxel selection for functional Magnetic Resonance Imaging (fMRI) brain data with the aim of finding a more complete set of probably correlated discriminative voxels, thus improving interpretation of the discovered potential biomarkers. The main difficulty in doing this is an extremely high dimensional voxel space and few training samples, resulting in unreliable feature selection. In order to deal with the difficulty, stability selection has received a great deal of attention lately, especially due to its finite sample control of false discoveries and transparent principle for choosing a proper amount of regularization. However, it fails to make explicit use of the correlation property or structural information of these discriminative features and leads to large false negative rates. In other words, many relevant but probably correlated discriminative voxels are missed. Thus, we propose a new variant on stability selection “randomized structural sparsity”, which incorporates the idea of structural sparsity. Numerical experiments demonstrate that our method can be superior in controlling for false negatives while also keeping the control of false positives inherited from stability selection.

Keywords: voxel selection, structural sparsity, stability selection,

*Corresponding Author

Email address: `chenhf@uestc.edu.cn` (Huaifu Chen)

1. Introduction

1.1. Problem Statement

Decoding neuroimaging data, also called brain reading, is a kind of pattern recognition that has led to impressive results, such as guessing which image a subject is looking at from his brain activity (Haxby et al., 2001), as well as in medical diagnosis, e.g., finding out whether a person is a healthy control or a patient. This pattern recognition typically consists of two important components: feature selection and classifier design. While the predictive or classification accuracy of these designed classifiers have received most attention in most existing literature, feature selection is an even more important goal in many practical applications including medical diagnosis where selected voxels can be used as biomarker candidates (Guyon and Elisseeff, 2003).

However, most traditional feature selection methods fail to discover in a stable manner the “complete” discriminative features accurately. They mainly aim to construct a concise classifier and they often select only a minimum subset of features, ignoring those correlated or redundant but informative features (Guyon and Elisseeff, 2003; Blum and Langley, 1997). In addition, the stability of the selected features is often ignored (Bühlmann and Van De Geer, 2011; Cover, 1965), because the inclusion of some noisy features or the exclusion of some informative features may not affect the prediction accuracy (Yu et al., 2008), which is their main objective. Therefore, a large number of uninformative, noisy voxels that do not carry useful information about the category label, could be included in the final feature detection results (Langs et al., 2011), while some informative, possibly redundant features might be missed.

In this paper, we focus on feature selection on functional MRI (fMRI) data where each voxel is considered as a feature. These features are often correlated or redundant. We focus on the “completeness” and “stability” of feature selection, i.e. aim to discover as many as possible informative but possibly redundant features accurately and stably, in contrast to most of the existing methods which mainly aim to find a subset of discriminative features which are expected to be uncorrelated. This way, potential biomarkers

revealed by the discovered discriminative voxels, in both cognitive tasks and medical diagnoses are expected to be more credible.

1.2. Advantages and Limitations of Sparse priors in Multivariate Neuroimaging Modeling

There are in general three main categories of supervised feature selection algorithms: filters, embedded methods, and wrappers (Guyon and Elisseeff, 2003). The filter methods usually separate feature selection from classifier development. For example, Fisher Score (Duda et al., 2000), is among the most representative algorithms in this category. The wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set, and the features are scored according to their predictive power. The embedded models perform feature selection during learning. In other words, they achieve model fitting and feature selection simultaneously. The following sparsity related feature selection models are all typical embedded methods, which we will mainly focus on in this paper.

In this paper, we consider commonly used supervised learning to identify the discriminative brain voxels from given training fMRI data. While the classification problem is considered most often, the regression problem can be treated in a similar way. We consider the following linear model.

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^{n \times 1}$ is the binary classification information and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the given training fMRI data and $\mathbf{w} \in \mathbb{R}^{p \times 1}$ is the unknown weights reflecting the degree of importance of each voxel. As a multivariate inverse inference problem, identification of discriminative voxels is based on the values of the weight vector \mathbf{w} and their importance is proportional to the absolute values of the components. Therefore, feature selection is also called support identification in this context, because the features corresponding to the nonzero \mathbf{w} components are considered as the relevant features.

Considering that the common challenge in this field is the curse of dimensionality $p \gg n$, we are focusing on sparsity-based voxel selection methods, because sparsity is motivated by the prior knowledge that the most discriminative voxels are only a small portion of the whole brain voxels (Yamashita et al., 2008).

However, sparsity alone is not sufficient for making reasonable and stable inferences. Plain sparse learning models often provide overly sparse

and hard-to-interpret solutions where the selected voxels are often scattered (Rasmussen et al., 2012), though they might be useful if a concise classifier is expected. Specifically, if there is a set of highly correlated features, then only a small portion of representative voxels are selected, resulting into a large false negative rate and a potential biomarker that is hard to trust. In addition, let us denote the support of the true sparse vector $\bar{\mathbf{w}}$ as S , and the number of its nonzeros as ℓ . For the success of finite sample recovery by the plain ℓ_1 norm regularized model, ℓ should be smaller than n . Let subsets of the columns of the design matrix \mathbf{X} larger than ℓ must be well conditioned. In particular, the design matrix \mathbf{X}_S should be sufficiently well conditioned and should not be too correlated to the columns of \mathbf{X} corresponding to the noisy subspace $\mathbf{X}_{\bar{S}}$ (Gaël Varoquaux, 2012).

Thus we have to extend the plain sparse learning model to incorporate important structural features of brain imaging data, such as brain segregation and integration, in order to achieve stable, reliable and interpretable results.

1.3. Existing Extensions of the Plain Sparse Model

As mentioned above, two common hypotheses have been made for fMRI data analysis: sparsity and compact structure. In sparsity, few relevant and highly discriminative voxels are implied in the classification task; in compact structure, relevant discriminative voxels are grouped into several distributed clusters, and the voxels within a cluster have similar behaviors and are, correspondingly, strongly correlated. Thus making use of these two hypotheses is very important, and we will review some state-of-the-art existing works in this direction.

Elastic net regression (Zou and Hastie, 2005) tries to make use of the voxel correlation by adding an ℓ_2 regularization, also called the Tikhonov regularization, to the classical ℓ_1 penalty (Ryali et al., 2012a) to deal with highly correlated features. Recently, other penalties have been added to consider the correlated features besides the Tikhonov regularization (Dubois et al., 2014). For example, both ℓ_1 penalization and Total-Variation (TV) penalization are used simultaneously for voxel selection (Gramfort et al., 2013), where the TV penalization is used to make use of the assumption that the activations are spatially correlated and the weights of the voxels are close to piece-wise constant. In addition, ℓ_2 -fusion penalty can be used if successive regression coefficients are known to vary slowly and can also be interpreted in terms of correlations between successive features in some cases (Hebiri and van de Geer, 2011). While these models based on both ℓ_1 norm

and other certain smoothing penalty, might achieve improved sensitivity over the plain ℓ_1 norm regularized model, they do not make use of any explicit prior grouping or other structural information of the features (Xia et al., 2010).

Correspondingly, another class of methods to make more explicit use of the segregation and integration of the brain, is based on structured sparsity models (Bach et al., 2012b; Schmidt et al., 2011; Chen et al., 2012), which have been proposed to extend the well-known plain ℓ_1 norm regularized models by enforcing more structured constraints on the solution. For example, the discriminative voxels are grouped together into few clusters (Baldassarre et al., 2012; Michel et al., 2011), where the (possibly overlapping) groups have often been known as a prior information (Xiang et al., 2012; Liu and Ye, 2010; Yuan et al., 2013; Jacob et al., 2009; Liu et al., 2009a; Ng and Abugharbieh, 2011). However, in many cases, the grouping information is not available beforehand, and one can use either the anatomical regions as an approximation (Batmanghelich et al., 2012), or the data driven methods to obtain the grouping information such as hierarchical agglomerative clustering (Ward hierarchical clustering, for example) and a top-down step to prune the generated tree of hierarchical clusters in order to obtain the grouping information (Michel et al., 2012; Jenatton et al., 2012).

While structural sparsity helps select the correlated discriminative voxels and is necessary for the “completeness” of the selected discriminative voxels, the result of feature selection may not be stable and is likely to include many noisy and uninformative voxels. For years, the idea of ensemble has been applied to reduce the variance of feature selection result (Hastie et al., 2009; Mota et al., 2014). Among them, one important class of methods for high dimensional data analysis is stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013). It is an effective way for voxel selection and structure estimation, based on subsamplings (bootstrapping would behave similarly). It aims to alleviate the disadvantage of the plain ℓ_1 norm regularized model, which either selected by chance non-informative regions, or even worse, neglected relevant regions that provide duplicate or redundant classification information (Mitchell et al., 2004; Li et al., 2012). This is due in part to the worrying instability and potential deceptiveness of the most informative voxel sets when information is non-local or distributed (Anderson and Oates, 2010; Poldrack, 2006). Correspondingly, one major advantage of stability selection is the control of false positives, i.e. it is able to obtain the selection probability threshold based on the theoretic

cal boundary on the expected number of false positives. In addition, stability selection is not very sensitive to the choice of the sparsity penalty parameter, and stability selection has been applied to the pattern recognition based on brain fMRI data and achieved better results than plain ℓ_1 norm regularized models (Ye et al., 2012; Cao et al., 2014; Ryali et al., 2012b; Mairal and Yu, 2013a; Meinshausen, 2013; Rondina et al., 2014). For example, SCoRS (Rondina et al., 2014) is an application of stability selection designed for the particular characteristics of neuroimaging data. Notice that we are focusing on the feature selection here. As for the prediction or classification accuracy, this ensemble or averaging idea has already been applied to reduce the prediction variance, and the examples include the bagging methods and forests of randomized trees (Breiman, 1996, 2001).

In order to make use of the assumption that these discriminative voxels are often spatially contiguous and result in distributed clusters, one proposed the idea of using common stability selection together with clustering (Gramfort et al., 2012; Gaël Varoquaux, 2012). Specifically, the clustering will be run after subsampling on training samples and random rescaling of features during each resampling of stability selection. The added clustering helps to improve the conditioning of resulted sub-matrices of the training data matrix. However, the random “rescaling” during their implemented stability selection is voxel-wise and fails to consider the spatial contiguity of the clustered discriminative voxels.

1.4. Our focus and contributions

In this paper, we propose a variant of stability selection based on structural sparsity, called “randomized structural sparsity”. It is implemented via the adoption of the “constrained block subsampling” technique for voxel-wise fMRI data analysis, in contrast to single voxel-wise subsampling in the classical stability selection. We expect it to achieve an improved sensitivity of the selected discriminative voxels. We show empirically that this “blocked” variant of stability selection can achieve significantly better sensitivity than alternatives, including the original stability selection, while keeping the control of false positives for voxel selection.

We need to point out that this new algorithm is beyond a simple summation of stability selection and structural stability. It has the following extra important advantage: in many cases where structural information such as clustering structures is only a rough approximation, i.e. neighboring voxels in the same brain area might be highly correlated though not necessarily all

informative, a.k.a. discriminative, the subsampling scheme can help remedy this via supervised refining and outlining of the true shapes of the discriminative regions, as showed by numerical experiments. Compared with Randomized Ward Logistic algorithm proposed in (Gramfort et al., 2012), our algorithm only needs to perform clustering once, and therefore is computationally more efficient.

The rest of the paper is organized as follows. In section 2, we introduce our new algorithm for stable voxel selection. In section 3, we demonstrate the advantages of our algorithm based on both synthetic data and real fMRI data in terms of higher sensitivity and specificity. In section 4, a short summary of our work and possible future research directions will be given.

2. The Proposed Method

2.1. Background and Motivation

Let us denote an fMRI data matrix as $\mathbf{X} \in \mathbb{R}^{n \times p}$ where n is the number of samples and p is the number of voxels with $n \ll p$, and corresponding classification information as $\mathbf{y} \in \mathbb{R}^{n \times 1}$. Here we consider only the binary classification and $\mathbf{y}_i \in \{1, -1\}$. While our main ideas can be applied to other models, we take the following sparse logistic regression for classification as our example to show the existing difficulties and our corresponding efforts, in detail.

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1 + \lambda \sum_{i=1}^n \log(1 + \exp(-\mathbf{y}_i(\mathbf{X}_i^T \mathbf{w} + c))) \quad (2)$$

where \mathbf{X}_i denotes the i -th row of $\mathbf{X} \in \mathbb{R}^{n \times p}$; $\mathbf{w} \in \mathbb{R}^{p \times 1}$ is the weight vector for the voxels and c is the intercept (scalar). The voxels corresponding to \mathbf{w}_i with large absolute value will be considered as the discriminative voxels.

Structured sparsity models beyond the plain ℓ_1 norm regularized models have been proposed to enforce more structured constraints on the solution (Bach et al., 2012b; Li et al., 2013; Mairal and Yu, 2013b), where the structure can be defined based on the feature correlation. As an important special case, the common way to make use of the clustering or grouping structure is to adopt the group sparsity induced norm (Bach et al., 2012a), as follows.

$$\min_{\mathbf{w}} \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2 + \lambda \sum_{i=1}^n \log(1 + \exp(-\mathbf{y}_i(\mathbf{w}^T \mathbf{X}_i + c))), \quad (3)$$

where \mathcal{G} is the grouping information. Compared with (2), the main difference is the regularization term; we are using a mixed ℓ_1/ℓ_2 norm. The model (3) belongs to the family of structural sparsity regularized feature selection models. The resulting penalty incorporating the parcellation information has been shown to improve the prediction performance and interpretability of the learned models, provided that the grouping structure is relevant (Yuan and Lin, 2006; Huang and Zhang, 2010; Jenatton et al., 2012; Bach et al., 2012b). In addition, the number of selected candidate features is allowed to be much larger when an additional group structure is incorporated, particularly when each group contains considerable redundant features (Jenatton et al., 2011; Xiang et al., 2015). Therefore, the parcellation is able to help improve the sensitivity of voxel selection (Flandin et al., 2002).

However, the group sparsity-induced norm regularized model (3) is expected to improve the sensitivity with respect to the plain ℓ_1 norm regularized model (2) due to the adopted mixed $\ell_{2,1}$ norm only if the grouping information \mathcal{G} is reliable enough. Obtaining an appropriate \mathcal{G} might be possible in practice from either the prior anatomical knowledge or data-driven methods based on the voxel correlation. However, many methods of obtaining \mathcal{G} are not incorporating the available classification or labelling information. Therefore, it is possible that only a subset of voxels in a certain group is discriminative. In such case, the model (3) often fails to make a segmentation, because it is likely to simultaneously choose all the voxels of a certain group or simultaneously choose none of them, due to the adoption of the ℓ_2 norm. In addition, just like the plain ℓ_1 norm regularized model, the difficulties of choosing a proper regularization parameter and lack of finite sample control of false positives still exist.

As mentioned above, an effective way to control the false positives and reduce the difficulty of choosing the proper regularization parameter when applying the sparsity regularization based models is stability selection (Meinshausen and Bühlmann, 2010), which has been applied for voxel selection or connection selection in brain image analysis (Rondina et al., 2014; Ye et al., 2012; Cao et al., 2014; Ryali et al., 2012b).

However, while the control of false positives can be achieved, a large false negative rate is often expected, especially in the case of redundant and correlated voxels, because this correlation prior is not explicitly taken into consideration.

2.2. The Key Component: Randomized Structural Sparsity

In this paper, we aim to stably identify the discriminative voxels including those probably correlated ones, for better interpretation of discovered potential biomarkers. To achieve this goal, we incorporate the spatial structural knowledge of voxels into the stability selection framework. The novelty of our research is to propose a “*randomized structural sparsity*”, which aims to integrate the stability selection and the common “structural sparsity”.

One important component of “randomized structural sparsity” is the subsampling based stability selection (Beinrucker et al., 2012), rather than the original reweighting-based stability selection (Meinshausen and Bühlmann, 2010). It has been shown that the former is likely to yield an improvement over the latter whenever the latter itself improves over a standalone pure ℓ_1 regularization model (Beinrucker et al., 2012). Moreover, subsampling is easier to extend to block subsampling and combine with structural sparsity.

Let us first briefly explain subsampling-based stability. For the training data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, subsampling based stability selection consists of applying the baseline, i.e. the pure ℓ_1 regularization model such as (2), to random submatrices of \mathbf{X} of size $[n/L] \times [p/V]$, where $[]$ is rounded off to the nearest integer number, and returning those features having the largest selection frequency. The original stability selection (Meinshausen and Bühlmann, 2010) can be roughly considered as a special case, where $L = 2$ and $V = 1$, except that the original stability selection (Meinshausen and Bühlmann, 2010) reweighs each feature (voxel, here) by a random weight uniformly sampled in $[\alpha, 1]$ where α is a positive number, and subsampling can be intuitively seen as a crude version of this by simply dropping out randomly a large part of the features (Beinrucker et al., 2012).

The other important component of randomized structural sparsity is to incorporate structural information, such as the parcelling information of the brain into consideration. The kind of partition information is based on either the prior anatomical knowledge of brain partition (Tzourio-Mazoyer et al., 2002), or the clustering results based on the fMRI data, as done in the structural sparsity model (3).

The above “randomized structural sparsity” is a general concept and might have different specific implementations in practice, depending on different data types and applications. For voxel-wise fMRI data analysis, we propose a specific implementation named “*constrained block subsampling*”, where by “constrained” we mean that the parcelling information will be respected to certain degree. The block subsampling (Lahiri, 1999) is adopted

because it is able to replicate the correlation by subsampling of blocks of data.

Specifically, each cluster $g \in \mathcal{G}$, consists of highly correlated voxels. After the block subsampling, the selected voxels from the same cluster will be considered as a group. In particular, the chosen voxels lying in a cluster $g \in \mathcal{G}$ are noted as a set $g' \subseteq g$. In addition, in order to make every brain partition, especially those of small sizes have a chance to be sampled during the block subsampling, we borrow some idea of “proportionate stratified sampling” (Särndal, 2003; de Vries, 1986), i.e. the same sampling fraction is used within each partition. The purpose is to reduce the false negatives, especially when the sizes of different partitions are of quite a range. Correspondingly, one can solve the following group-sparsity based recovery model.

$$\min_{\mathbf{w}'} \sum_{g' \subseteq g \in \mathcal{G}} \|\mathbf{w}'_{g'}\|_2 + \lambda \sum_{i \in \mathcal{J}} \log(1 + \exp(-\mathbf{y}_i(\mathbf{w}'^T \mathbf{X}'_i + c))) \quad (4)$$

where \mathbf{w}' and \mathbf{X}' are corresponding parts of \mathbf{w} and \mathbf{X} , respectively, based on the selected voxels during the subsampling, and \mathcal{G} is a predefined partitions of the brain, based on the either biological knowledge or data driven learning or estimation such as clustering. \mathcal{J} is the set of the indices of the selected samples during the current subsampling.

Notice that while “constrained block subsamplings” respects the prior knowledge \mathcal{G} , it also provides the flexibility that the resulting discriminative regions can be of any shape, and the final selected voxels of each cluster can be only a portion of all of it, because the subsampling makes the selection frequency score be able to outline shapes of the true discriminative regions, whose sizes may not be exactly the same as the sizes of the original partitions defined by \mathcal{G} . This kind of flexibility is important because the neighboring voxels belonging to the same brain area are not necessarily all significantly discriminative voxels, though they might be highly correlated. In other words, we aim to seek sets of correlated voxels with similar associations with the response (or labels), if only part of but not all of the correlated features have a similar association with the response, as mentioned in (Witten et al., 2014).

Furthermore, for our case of small samples and very high dimensional feature space, we need to consider the bias-variance dilemma or bias-variance tradeoff (Geman et al., 1992). In general, we would like to pay a little bias to save a lot of variance, and dimensionality reduction can decrease variance by simplifying models (James et al., 2013). Correspondingly, while we can

still use the (4) as the baseline subproblem for our stability selection framework, we prefer a simple “averaging” idea (Gaël Varoquaux, 2012) applied to (4), because (Park et al., 2007) has showed that when the variables or features were positively correlated, their average was a strong feature, and this yielded a fit with lower variance than the individual variables. Specifically, by averaging the voxels picked by the block subsampling lying in the same group as a single super-voxel, the model (4), can be further reduced to the following low dimensional version

$$\min_{\tilde{\mathbf{w}}} \sum_{g' \subset g \in \mathcal{G}} |\tilde{\mathbf{w}}_{g'}| + \lambda \sum_{i \in \mathcal{I}} \log(1 + \exp(-y_i(\tilde{\mathbf{w}}^T \tilde{\mathbf{X}}_i + c))) \quad (5)$$

where $\tilde{\mathbf{w}} \in \mathbb{R}^q$, and q is the number of clusters. $\tilde{\mathbf{w}}_{g'}$ is an average of voxels in the subset g' of cluster $g \in \mathcal{G}$, and $\tilde{\mathbf{X}} \in \mathbb{R}^{[\alpha n] \times p}$ is the corresponding averaged \mathbf{X} . Thus the number of variables in the sparse recovery model (5) is greatly reduced to the number of clusters. This way, the resulted recovery problem (5) is of much smaller scale and therefore can be solved quite efficiently. In addition, the properties of the resulting new data matrix $\tilde{\mathbf{X}}$ is greatly improved due to de-correlation via the clustering of correlated columns. The analysis of a better-posed compatibility constant for the $\tilde{\mathbf{X}}$ was proposed in (Bühlmann et al., 2013). The idea of averaging, also called feature agglomeration (Flandin et al., 2002), was also applied in (Gramfort et al., 2012). If the j -th column of $\tilde{\mathbf{X}}$ is selected due to the large magnitude of $\tilde{\mathbf{w}}_j$, then its represented picked blocked voxels lying in the group $g^{(j)} \in \mathcal{G}$ ($j = 1, 2, \dots, q$) of \mathbf{X} are all counted to be selected, in the non-clustered space. Its corresponding score \mathbf{s}_i will be updated ($i = 1, 2, \dots, p$). Notice that the averaging of subsamplings is more than a simple spatial smoothing, due to different subsampling results of different stability selection iterations. Therefore, the boundaries of the detected discriminative regions can be trusted to certain accuracy.

2.3. Algorithmic framework

We first obtain the structural information about the brain. Here we perform a data-driven clustering operation to partition the voxels into many patches according to their strong local correlations. In our algorithm, both the common K-means and the spatially constrained spectral clustering algorithm (Craddock et al., 2013) implemented written as a Python software (http://www.nitrc.org/projects/cluster_roi/) are used in our experiments. We denote the set of the groups via the clustering algorithm as \mathcal{G} ,

whose cardinality is denoted as q , which is usually much less than p and comparable with n . Notice that in (5), the number of unknowns is reduced from p to the number of clusters, i.e. q . While the number of samples is a fraction of the total samples, for example, $\lfloor n/2 \rfloor$. In this paper, we typically choose the q at least 2 times larger than the number of samples but smaller than 5 times of the number of samples in practice.

Next comes the “constrained block subsamplings”. Denote the number of resamplings as K . This blocked variant of stability selection is different from the classical stability selection in terms of the subsampling on the features, i.e. the columns of the data matrix \mathbf{X} . But it shares the same way as the classical stability selection when performing subsampling of the observations, i.e. the rows of the data matrix \mathbf{X} . Let the subsampling fraction be $\alpha \in [0, 1]$ and let \mathcal{J} denote the indices of selected rows and the cardinality of \mathcal{J} is $\lfloor \alpha n \rfloor$, where $\lfloor \cdot \rfloor$ is rounded off to the nearest integer number. Then “constrained block subsamplings” are applied to the voxels, i.e. the columns of \mathbf{X} as mentioned in last section. Notice that our algorithm only runs the clustering once and the following “constrained block subsamplings” resulted in a much smaller size of ℓ_1 problem, where the number of unknowns is equal to the number of clusters. Therefore, our algorithm is not computationally expensive.

The procedure of our algorithm is summarized below.

The Algorithmic Framework of Constrained Blocked Stability Selection Method:

Inputs:

- (1) Datasets $\mathbf{X} \in \mathbb{R}^{n \times p}$
- (2) Label or classification information $\mathbf{y} \in \mathbb{R}^n$
- (3) Sparse penalization parameter $\lambda > 0$
- (4) Number of randomizations K for each stage; subsampling fraction $\alpha \in [0, 1]$ in terms of rows of \mathbf{X} ; subsampling fraction $\beta \in [0, 1]$ in terms of columns of \mathbf{X} ;
- (5) Initialized stability scores: $\mathbf{s}_i = 0$. ($i = 1, 2, \dots, p$)

Output: Stability scores \mathbf{s}_i for each voxel. ($i = 1, 2, \dots, p$)

Obtain a brain parcellation. For example, perform the clustering of voxels based on their spatial correlation and denote the number of clusters as q

for $k=1$ to K

1: Perform sub-sampling in terms of rows: $X \leftarrow X_{[\mathcal{J},:]}, y \leftarrow y_{\mathcal{J}}$ where $\mathcal{J} \subset \{1, 2, \dots, n\}$, $\text{card}(\mathcal{J}) = [\alpha n]$, the updated $X \in R^{[\alpha n] \times p}$, and the updated $\mathbf{y} \in R^{[\alpha n]}$.

2: Perform constrained block subsampling in terms of columns (voxels): $X' \leftarrow X_{[:,\mathcal{I}]}$, where $\mathcal{I} \subset \{1, 2, \dots, p\}$, and $\text{card}(\mathcal{I}) = [\beta p]$

3: Use the current clustering, and calculate the mean of randomly picked voxels within each cluster: $\tilde{X} \leftarrow \text{mean}(X')$, $\tilde{X} \in R^{\alpha n \times q}$

4: Estimate $\tilde{\mathbf{w}} \in \mathbb{R}^q$ from \tilde{X} and \mathbf{y} with sparse logistic regression (5).

5: Set weights for the randomly picked voxels with estimated coefficients of the averaged voxels: $\mathbf{w}^{(k)} \leftarrow \tilde{\mathbf{w}}$, $\mathbf{w}^{(k)} \in \mathbb{R}^{[\beta p]}$

6: $\mathbf{s}_i = \mathbf{s}_i + 1$, if $i \in \text{supp}(\mathbf{w}^{(k)})$, for $i = 1, 2, \dots, p$.

end for

2.4. Some Preliminary Rethinking of Our Algorithms

Basically, the original stability selection proposed in (Meinshausen and Bühlmann, 2010) is mainly on random subsampling of observations, i.e. the rows of \mathbf{X} . As the paper by (Beinrucker et al., 2015) has also pointed out, the random subsampling in terms of observations can in general guarantee the finite control of false positives, even though different base methods are adopted. Therefore, while we are using a more complicated base method (5) than the plain ℓ_1 norm regularized model, the finite control of false positives can be still achieved. However, the corresponding new theoretical result in terms of bounding the ratio of the expected number of false positive selections over

the total number of features (false positive rate) needs to be addressed in the future work.

It is natural that we adopt the structural sparsity regularized models such as (5), as the base methods of stability selection. As (Bach et al., 2012b,a; Flandin et al., 2002) pointed out, the regularization term incorporating the parcellation information has been shown to improve the interpretability of the learned models and the detection sensitivity of voxel selection for the functional MRI data, provided that the parcellation information is quite relevant.

However, the parcellation information might be not very accurate. Any fixed brain parcellation indeed might bring certain degree of bias or arbitrariness. In this paper, we turn to help of the block subsamplings. While we present some intuitive explanation in Section 2.2, a thorough study of the effect of block subsampling on reducing the arbitrariness is not presented in this paper and constitutes an important future research topic.

Here we need to point out that the method proposed in (Gaël Varoquaux, 2012) does not suffer the bias caused by a fixed parcellation, because the clustering is performed on each step of stability selection after the randomized rescaling on each feature. However, from the computational point of view, our adopted onetime parcellation helps improve the computational efficiency, because clustering takes a large proportion of running times of both our algorithm and the algorithm by (Gaël Varoquaux, 2012). Some preliminary comparison of running time of different algorithms are presented in Section 3.6.

3. Numerical Experiments

In this paper, we compare our algorithm with the classical univariate voxel selection method, and with state-of-the art multi-voxel pattern recognition methods, including T-test, ℓ_2 -SVM, ℓ_2 Logistic Regression, ℓ_1 -SVM, ℓ_1 Logistic Regression, randomized ℓ_1 logistic regression, Smooth Lasso (Hebiri and van de Geer, 2011) and TV-L1 (Gramfort et al., 2013) and Randomized Ward Logistic (Gaël Varoquaux, 2012). Here randomized ℓ_1 logistic regression is based on the original stability selection (Meinshausen and Bühlmann, 2010) and random reweighing on the features.

The T-test is implemented as an internal function in MATLAB. ℓ_2 -SVM, ℓ_2 Logistic Regression, ℓ_1 -SVM, and ℓ_1 Logistic Regression, have been implemented in LIBLINEAR (Fan et al., 2008) or SLEP (Sparse Learning with

Efficient Projections) software (Liu et al., 2009b). Randomized ℓ_1 logistic regression is written based on the available ℓ_1 logistic regression code. TV-L1 and Randomized Ward Logistic are implemented in Python, and integrated into NiLearn, a great Python software for NeuroImaging analysis <http://nilearn.github.io/index.html>. We were kindly provided with the source code by their developers. For the hyper-parameters such as the regularization parameters, their choices are mostly based on cross validation unless specified otherwise.

3.1. Settings of Algorithms

For our algorithm, the block size might affect the performance of our algorithm (Lahiri, 2001). Given the number of blocks, there is an inherent trade-off in the choice of block size. When only a very limited number of randomizations are allowed, big blocks will most likely not match the geometry of the true support and easily result in many false positives. But the condition of too small blocks is likely to result in many false negatives due to the likely ignorance of the local correlations of neighboring voxels. The block size is not optimized in the following experiments via the probable prior knowledge of the discriminative regions, but it still achieves an impressive performance. It was set to be 3×3 in synthetic data and $4 \times 4 \times 4$ in the real fMRI data experiment, respectively. We set the subsampling rate $\alpha = 0.5$ and $\beta = 0.1$. For our synthetic data and the real fMRI data, we set the total resamplings $K = 50$ and $K = 200$, respectively. The resampling number of random ℓ_1 logistic regression is 500 in all of our experiments. The choice of the number of resamplings is only empirical here.

3.2. Evaluation Criteria

We would like to demonstrate that our method can achieve better control of false positives and false negatives than alternative methods, due to our incorporation of both stability selection and structural sparsity.

For the synthetic data, we can directly use the precision-recall curve since we know the true discriminative features. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. We also plot the first T discriminative voxels discovered by different algorithms, where T is the number of true discriminative features. It provides a snapshot to display how many noisy features are included in the selected features of different algorithms.

For real fMRI data, we first show brain maps obtained by the feature weights, which are not thresholded for visualization purposes, meaning that the zeros obtained are actually zeros. We are able to determine the discriminative regions, revealed by different algorithms, based on our visual inspection.

In addition to vision inspection and experience, we would also like to find an objective threshold. In general, voxels whose corresponding weights have larger magnitude than this threshold will be considered as the discriminative voxels and will be shown in the brain maps. However, the setting of a threshold value is quite difficult and may adopt different schemes in different situations. While a through study of threshold setting is out of reach of this paper, we consider to use the cross-validation based on prediction accuracy and find out the threshold value corresponding to the highest prediction accuracy. However, as (Hofner et al., 2014) has pointed out that the prediction accuracy and variable selection are two different goals. Different features might result into the same level of classification accuracy. Therefore, it is often acceptable to develop some heuristic for setting the threshold value for feature selection, beyond cross validation, such as the method by (Fellinghauer et al., 2013). Another kind of important method to set the threshold value is based on FDR control with multivariate p-values (Chi et al., 2008).

The main feature of our algorithm is its improved sensitivity while maintaining good specificity. In order to prove that the extra probably discriminative regions discovered only by our algorithm are true and stable positives, we adopt the following two methods. One is to take these extra selected voxels to construct a classifier to perform classification on the test data. A satisfying classification accuracy can at least prove the existence of true positiveness. Notice that, while we have mentioned that the prediction accuracy is not very reliable criterion for model selection, high prediction accuracy can in generally tell us that at least portion of these voxels are truly discriminative. If a high prediction accuracy is achieved, we are likely to believe that the corresponding brain regions are discriminative, though their sizes might be not accurate. The other is to perform a false positive estimation scheme based on a permutation test in order to calculate the ratio of false positives among all the finally selected voxels (Rondina et al., 2014).

3.3. Synthetic Data

We simulated simple case control analysis model and work on $46 \times 55 \times 46$ brain images including 27884 voxels of interest. We generated 50 observations for each group, i.e. the control group and the case (patient) group.

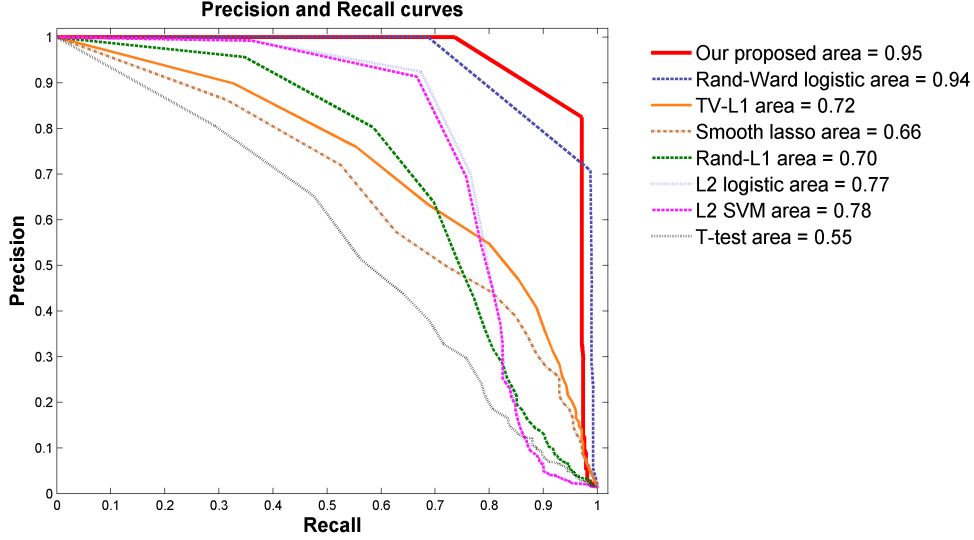


Figure 1: Precision-Recall Curve on the new Synthetic Data: Our method can achieve the best control of both false positives and false negatives, as well as the largest AUC (Area Under Curve) value .

There were five discriminated clustered features with 383 total voxels in the frontal lobe, parietal lobe, occipital lobe, and subcortical regions. Each cluster contained more than 30 voxels, as showed in Figure 2 in yellow.

The elements in the first two clustered features: $x(k)_{i,n} = k + \epsilon(k)_{i,n}$ (case), $y(k)_{j,n} = \eta(k)_{j,n}$ (control), where $i, j = 1, 2, \dots, 50$ representing the index of persons of each group, and $k = 1, 2$ representing the index of the first two clustered features, and $n = 1, 2, \dots, 100$ representing the index of features of each cluster. $\epsilon(k)_{i,n}$ and $\eta(k)_{j,n}$ are Gaussian i.i.d distributed. The elements in the other three clustered features are spatially distributed patterns, which are Gaussian i.i.d distributed and constrained by linear condition: $x(k)_{i,n} = \epsilon(k)_{i,n}$, $x(k)_{j,n} = \eta(k)_{j,n}$ $\sum_{k=3}^5 x(k)_{i,n} > 1$ (case), and $\sum_{k=3}^5 y(k)_{j,n} < 1$ (control), where $i, j = 1, 2, \dots, 50$ representing the index of persons of each group, and $k = 3, 4, 5$ representing the index of the last

three clustered features, and $n = 1, 2, \dots, 100$ representing the index of features of each cluster. As above, $\epsilon(k)_{i,n}$ and $\eta(k)_{j,n}$ are also Gaussian i.i.d distributed. The features were spatially clustered in different brain regions. We also simulated the other voxels in whole brain image X as Gaussian noise. Notice that these are distributed multivariate discriminative patterns, each of which consists of 3 voxels from each of the last 3 clusters, respectively. For the clustering algorithm used in our algorithm, we use K-means and the number of clustering is equal to 200.

We would like to show that our method can achieve both accuracy and completeness in terms of discovery of discriminative features. Here accuracy means a small false positive rate and completeness means a small false negative rate. In Figure 1, we use Precision-Recall Curve to demonstrate this advantage of our method. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. While still keeping good control of false positives, our algorithm, together with Randomized Ward Logistic are the most sensitive, i.e. discovering the almost “complete” set of discriminative features. Notice that the standard stability selection, i.e. randomized ℓ_1 algorithm does not work well in this case.

In Figure 2, we plot out the identified discriminative voxels corresponding to the top 383 weights of largest magnitude of these different involved algorithms, where 383 is the number of true discriminative voxels displayed in the subplot of the most upper-right corner. Our algorithm, together with the Randomized Ward Logistic algorithm, discover more clustered true positive features than others. Moreover, our algorithm is computationally more efficient than Randomized Ward Logistic because we run the clustering only once. Notice that the synthetic data is 3-D and therefore is hard to visualize in 2-D. So the performance comparison of each algorithm is more directly displaying via the Precision-Recall curve of Figure 1. Nevertheless, Figure 2 can be a useful supplement, as an illustration of the performance of the achieved sensitivity and specificity of different voxel-selection algorithms.

3.4. Real fMRI Data- Chess-Master Data

In this experiment, we aim to identify the brain activation pattern of a Chinese-chess problem-solving task in professional Chinese-chess grandmasters. 14 masters on Chinese chess were recruited and studied. All subjects

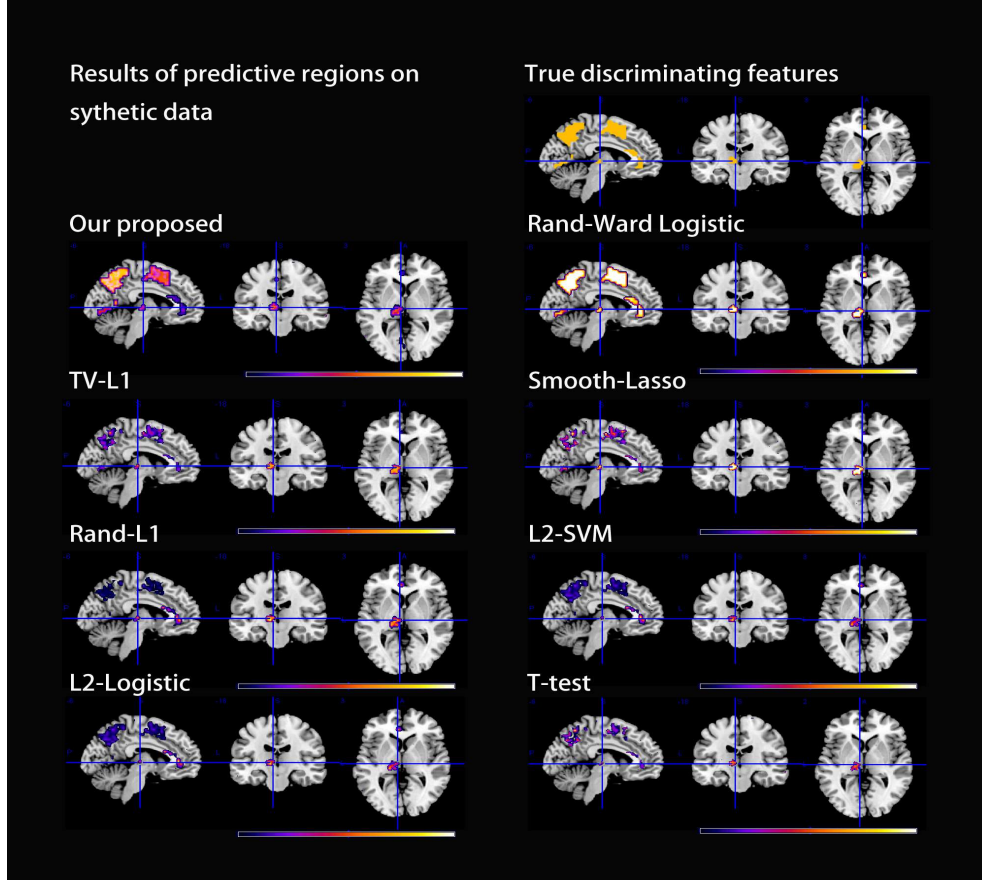


Figure 2: Result on the Synthetic Data: Our algorithm together with Randomized Ward Logistic algorithm can find more true discriminative voxels with very few noisy voxels.

were right-handed and had no history of psychiatric or neurological disorder. During the fMRI scanning, subjects were presented with two kinds of stimuli: a blank chessboard and patterns of Chinese chess spot game with checkmate problems. Each condition was presented for 20s, with a 2s-long break between. The block was repeated nine times with different problems in each block. The break between each block is also 2s. There were 9 blocks overall. In consideration of the delay of Blood Oxygenation Level Dependent (BOLD) effect (Aguirre et al., 1998) and the hypothesis that the master may solve the problem in less than 20s, we selected the 4th-8th images of each state in each block for classification. That is, the number of observations of each subject for classification is 90, among which 45 are in blank states

while the other 45 are in task states. We used an averaged data from all 14 grandmasters. Data Acquisition and Preprocessing Scanning was performed on a 3T Siemens Trio system at the MR Research Center of West China Hospital of Sichuan University, Chengdu, China. T2-weighted fMRI images were obtained via a gradient-echo echo-planar pulse sequence (TR, 2000ms; TE, 30ms; flip angle=90°; whole head; 30 axial slice, each 5mm thick (without gap); voxel size= $3.75 \times 3.75 \times 5mm^3$). fMRI images were pre-processed using Statistical Parametric Mapping-8 (SPM8, Wellcome Trust Centre for Neuroimaging, London, UK. <http://www.fil.ion.ucl.ac.uk/spm>). Spatial transformation, which included realignment and normalization, was performed using three-dimensional rigid body registration for head motion. The realigned images were spatially normalized into a standard stereotaxic space at $2 \times 2 \times 2 mm^3$, using the Montreal Neurological Institute (MNI) echo-planar imaging (EPI) template. A spatial smoothing filter was employed for each brain's three-dimensional volume by convolution with an isotropic Gaussian kernel (FWHM= 8 mm) to increase the MR signal-to-noise ratio. Then, for the fMRI time series of the task condition, a high-pass filter with a cut-off of 1/128 Hz was used to remove low-frequency noise. Among all 90 fMRI samples, each of them was of size $91 \times 109 \times 91$. For the clustering algorithm used in our algorithm, we use K-means and the number of clusters is equal to 200.

Figure 3 shows brain maps based on the weights or scores of voxels of different algorithms. The scores are not thresholded for visualization purposes, meaning that the zeros obtained are actually zeros. One can observe that despite being fairly noisy, the most significant localized discriminative regions of the brain, identified by different algorithms, can be visually recognized. Even with the same number of selected voxels, our method is expected to achieve the best balance of controlling of both the false positives and false negatives.

In general, when identifying the potential biomarkers, controlling the false positives should be the first priority. They need to be treated carefully and controlled strictly. So we need to set a threshold value to filter out at least the apparent noisy features, which are either too scattered or in the wrong regions from the existing confirmed knowledge. We carefully set the threshold values for results of different algorithms in order to control the false positives and obtain a cleaner brain map. Notice that in this experiment, a common threshold-setting method based on the prediction accuracy via cross validation does not work well, because this is a very simple cognitive task and the

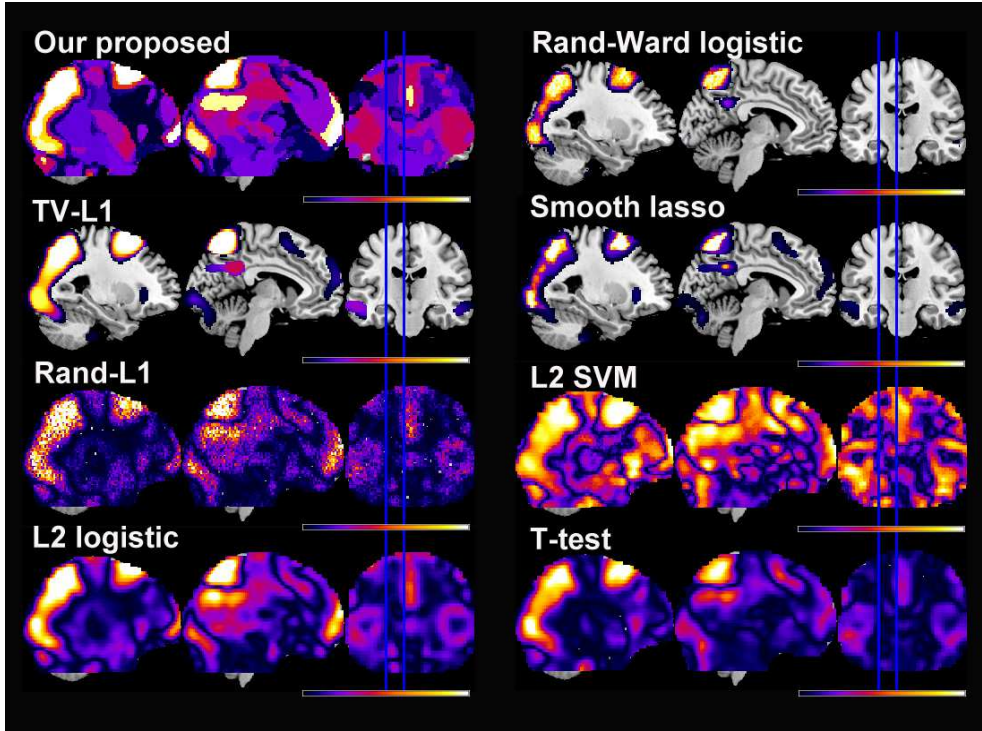


Figure 3: Score maps (unthresholded) estimated by different methods. The most significant areas discovered by our method are quite spatially contiguous and are of high contrast with other areas.

involvement of many noisy features or using only a small number of true positives can also achieve nearly 100% accuracy. Figure 4 is the result after thresholding out the apparent noisy features as either too scattered or in the unreasonable area. The false positives are expected to be well controlled. We can see that our algorithm is most sensitive and identified several extra brain regions. We construct classifiers based on each of these extra regions and test their predictive power. They are more than 95% accurate, so these extra regions are very likely to be relevant.

We take a further look at the result of Figure 4 from the viewpoint of brain science. All the multivariate pattern feature selection methods successfully identified at least partial task-related prefrontal and parietal and occipital lobe regions. These results indicate a co-working pattern of the cognitive network and default mode network of the human brain during our board game task state. However, compared with most alternative algorithms

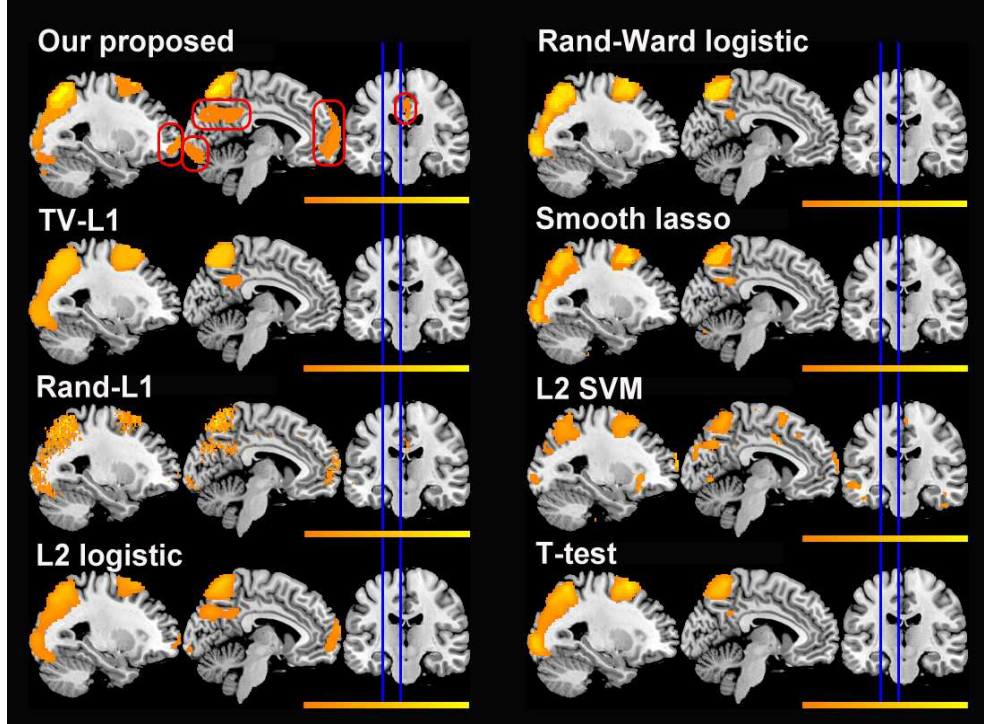


Figure 4: Score maps (thresholded) estimated by different methods. Our algorithm reveals more predictive areas.

besides the common stability selection, our proposed method identifies much more brain regions in the medial prefrontal cortex and precuneus gyrus that are functional and structural central hubs in the default mode network and in occipital lobe which contains parts of visual cortex. The common stability selection, i.e. randomized ℓ_1 logistic regression is able to identify the medial prefrontal gyrus, but it misses the precuneus. Moreover, the common stability selection is likely to return a result that is slightly more scattered, which does not match the second hypothesis about continuousness and compactness. Of even greater concern is that fact that its scattered results make it difficult to distinguish between true positives from false positives. In addition, common stability selection required many more subsamplings, for example, 500 times here compared with our method which only takes 50 subsamplings. This result verifies one of the main advantages of our method, namely its computational efficiency, which is especially important for high dimensional problems. In section 3.6, we will present the running time com-

parison of different algorithms. Our method also has better inference quality due to the incorporation of prior structural information of the fMRI data. As mentioned before, this computational efficiency also comes from the even smaller size of the subproblem (5) due to the adoption of the averaging idea within a cluster.

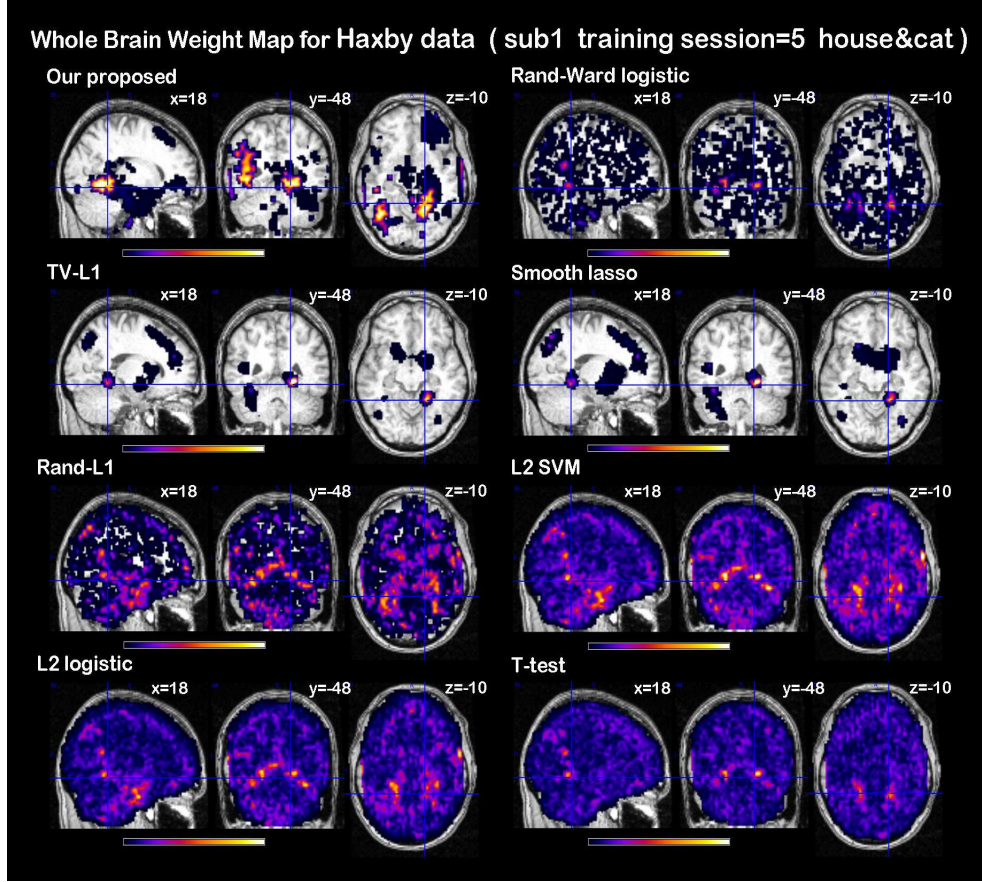


Figure 5: Score maps (unthresholded) as estimated by different methods on fMRI datasets (Cat vs House) using the first 5 sessions of training data. Despite being fairly noisy, located discriminative brain regions by different algorithm are well highlighted.

3.5. Real fMRI Data- Haxby Cognitive Task Data

We also test our algorithm on a public, block-design fMRI dataset from a study on face and object representation in human ventral temporal cortex (Haxby et al., 2001). The set, which can be downloaded at <http://data.pymvpa.org/datasets/haxby>



Figure 6: Score maps as estimated by different methods on fMRI datasets (Cat vs House) using first 5 sessions of training data. The threshold is determined based on cross-validation for the highest prediction accuracy. Our algorithm can achieve the best performance by finding a larger number of true discriminative voxels than alternatives and keeping the false positives into a very low level. Most of the alternatives have a large number of false positives, except the Randomized Ward Logistic method, which however, only finds a very small number of true discriminative voxels, although its estimated false positives is 0.

consists of 6 subjects with 12 runs per subject. In each run, the subjects passively viewed grayscale images of eight object categories, grouped in 24s blocks separated by rest periods. Each image was shown for 500ms and was followed by a 1500ms inter-stimulus interval. Full-brain fMRI data were recorded with a volume repetition time of 2.5s. Then a stimulus block was covered by roughly 9 volumes. For a complete description of the experimen-

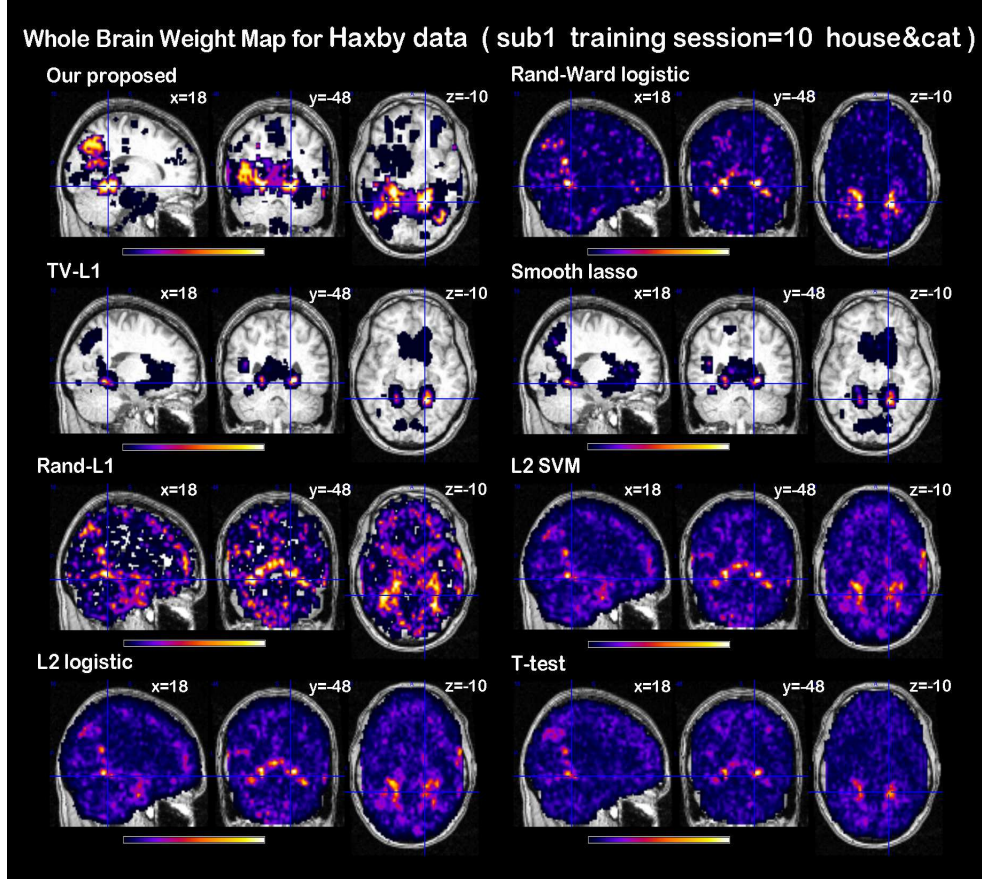


Figure 7: Score maps (unthresholded) as estimated by different methods on fMRI datasets (Cat vs House) using the first 10 sessions of training data. Despite being fairly noisy, located discriminative brain regions by different algorithm are well visually recognized.

tal design, fMRI acquisition parameters, and previously obtained results see the reference (Haxby et al., 2001). There is no smoothing operation on this data. In this paper, we consider the fMRI data of the first subject when classifying the “house” and “cat”, which consists of 12 sessions in total. The number of samples, for the first 5 sessions, is 90. The number of training samples evenly increases to 180 when the number of sessions are 10. We adopt the spatially constrained spectral clustering algorithm, implemented in a python software “PyClusterROI” (Craddock et al., 2013). The number of clustering is 200 when the number of sessions is 5 and evenly increases to 400 when the number of sessions is 10.

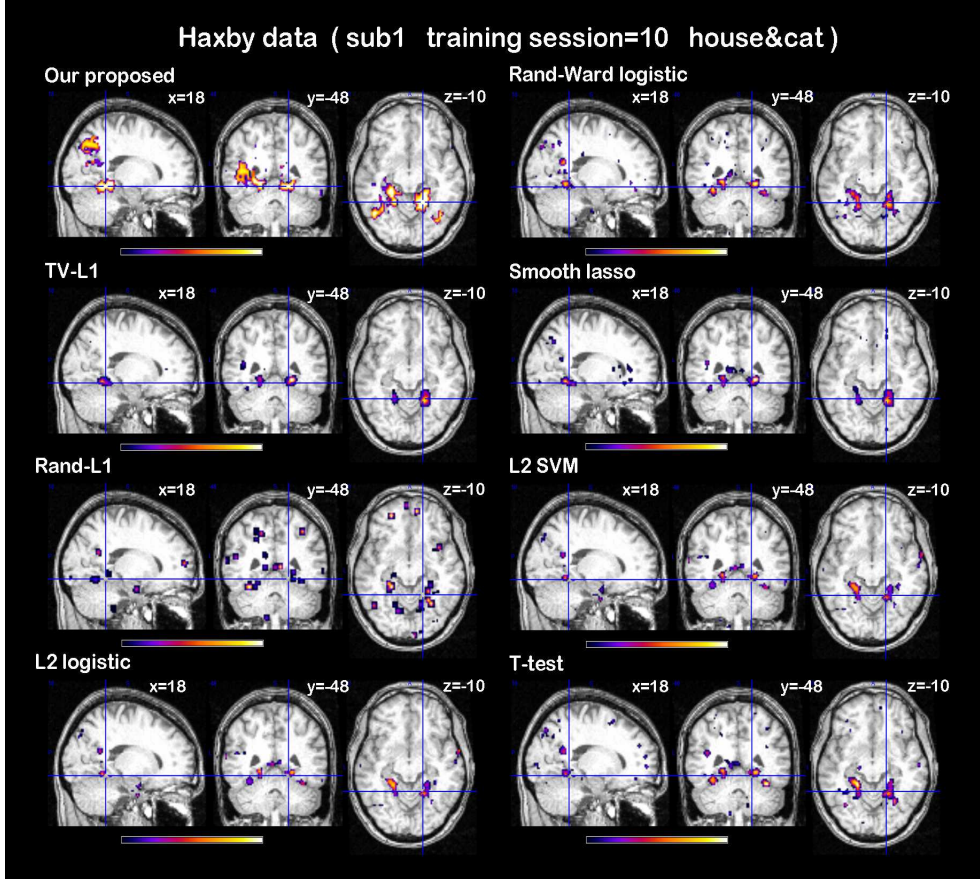


Figure 8: Score maps as estimated by different methods on fMRI datasets (Cat vs House) using first 10 sessions of training data. The threshold is determined based on cross-validation for the highest prediction accuracy. Our algorithm can achieve the best performance by finding a large number of true discriminative voxels than alternatives and keeping the false positives into a very low level. Most of the alternatives have a larger number of false positives, except the Randomized Ward Logistic method, which however, only finds a very small number of true discriminative voxels, although its estimated false positives is 0.

We use the first T sessions as training samples to perform the feature selection, where $T = 5, 6, 7, 8, 9, 10$. Then we obtain a prediction score of these selected features, on $12 - T$ remaining sessions, which are used as the test samples. Due to the limited length of this paper, we only show the brain maps obtained when we use the first 5 and 10 sessions as the training data.

Figures 5 and 7 are brain maps based on the scores of different meth-

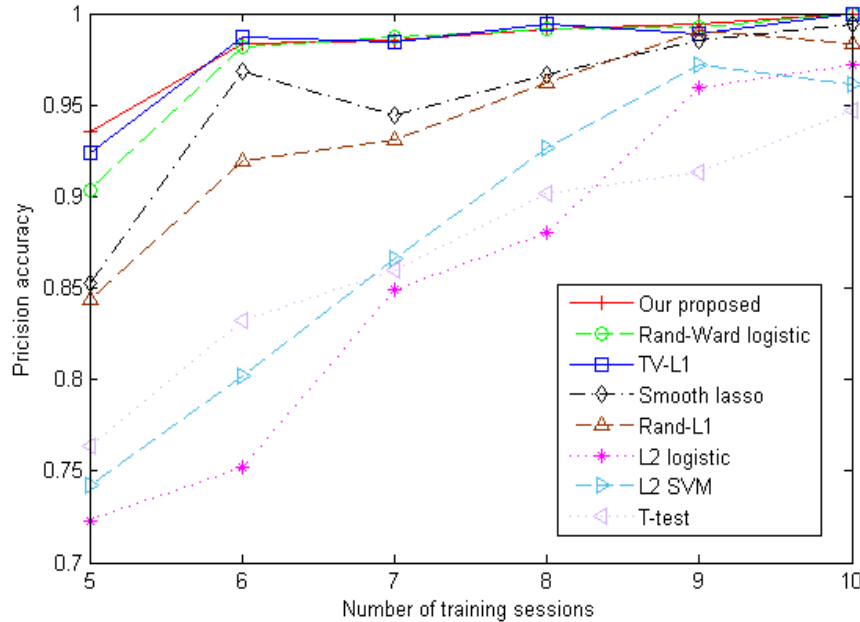


Figure 9: The classification accuracies of different algorithms when classifying the Cat and House. Our algorithm is among those achieving the highest prediction accuracy in general.

ods: the first 5 and 10 sessions, respectively. The scores are not thresholded for visualization purposes. Figures 6 and 8 show the thresholded maps of different algorithms, when we use the first 5 and 10 sessions as the training data, respectively. The threshold values of different algorithms can be different. The T-test is based on the common requirement that p-value is less than 0.001 for FDR control. All the rest of the methods are based on cross-validation, where a linear ℓ_2 -SVM classifier is used and the remaining 7 or 2 sessions are used as test data. The candidate threshold value corresponding to its own best prediction accuracy for each algorithm is chosen. Specifically, for our algorithm, we use almost the same threshold-setting procedure as the Python code of the Randomized Ward Logistic algorithm by (Gaël Varoquaux, 2012). Basically, we first rank the voxels according to the selection scores and set the thresholds from 0.3 to 0.9 with step size of 0.1 and select the final threshold value corresponding to the high classification accuracy on the testing set (the remaining sessions except those for training).

We can see that in general, our algorithm is among the most sensitive

Algorithm	Rand Ward Logistic	TV-L1	Smooth Lasso	Rand L1
Pred. Accu.	97.22%	100%	100%	97.22%
Algorithm	L2-SVM	L2-Logistic	T-test	
Pred. Accu.	94.44%	94.44%	94.44%	

Table 1: The prediction (classification) accuracy of the classifier on the voxels selected by RSS subtract those selected by other methods, respectively. Here the first 10 sessions are training data, and the rest 2 sessions are test data.

algorithms. In order to test whether the extra voxels selected only by our algorithm are of convincing prediction power (necessarily for potentially the true positives), we build an ℓ_2 logistic regression classification model based on these extra voxels (i.e., the discriminative voxels obtained by our proposed method subtracts those selected by other methods, respectively). The prediction accuracy of the resulted classifier are listed below in Table 1. We can see that the extra voxels selected by our method give high classification accuracy, showing that at least part of these extra selected voxels could be the relevant voxels of the task. In the following paragraph, we try to explain and validate the trueness of the discovered discriminative voxels from the viewpoint of neuroscience.

Our results of threshold and unthreshold maps both show the same phenomenon that was described in the original case study (Haxby et al., 2001), where the PPA and FFA are included. The area of mean response regions across all categories, selected by (Haxby et al., 2001), is in consistency with the common regions within our different cluster settings in the axial view. In the unthresholded mapping, the contours are quite similar with different numbers of clusters in the leftmost column; in the thresholded mapping, selected features are near the same positions.

As a variant of stability selection, our algorithm maintains the finite sample control of false positives. Our algorithm’ advantage is its improved sensitivity of feature selection comparing to other alternatives. Since there is no ground truth for evaluation, we have tested whether our detected voxels or regions by various algorithms are stable and unlikely to be false positives. We did this by adopting the false positive estimate scheme used in (Rondina et al., 2014), which is based on a permutation test and cross validation. In Figure 8, the result of our algorithm shows a selection of 1247 voxels with only 23 likely to be false positives. While TV-L1 and Smooth Lasso found larger number of discriminative voxels, 2611 and 1804, respec-

Algorithm	Ours	Rand Ward Logistic	TV-L1	Smooth Lasso
No.Selected	1247	116	2611	1804
No.False Positives	23	0	881	977
Algorithm	Rand L1	L2-SVM	L2-Logistic	T-test
No.Selected	1333	1499	1549	1045
No.False Positives	224	204	198	151

Table 2: “No. Selected” means the number of selected voxels after thresholding when using the first 10 sessions. “No.False Positives” is the number of probable false positives among all the selected voxels, estimated via permutation test and cross-validation, as suggested in (Rondina et al., 2014).

tively, and they also have 881 and 977 voxels, respectively, that are likely to be false positives. Even the original stability selection, i.e., Rand L1, has 224 estimated false positives among its selected 1333 discriminative voxels. L2-SVM and L2-Logistic also have around 200 estimated false positives. T-test has over 150 false positives. While our algorithm shares some common components with the Randomized Ward Logistic algorithm, the results are quite different. Randomized Ward Logistic method is more conservative in terms of controlling false positives, at least in its default settings. Its selected voxels have no false positives by the false positive estimate scheme. However, it only reveals 116 discriminative voxels. Its conservation in this case can even be observed from the unthresholded Figures 7. In contrast, our algorithm finds a large number of true discriminative voxels and keeps the false positives into a very low level. The summary of this result is in Table 2.

Now we look at the predictive power of the best selected voxels of different algorithms. These prediction results are reported in Figure 9. We consider to pick T sessions as the training data and the remaining 12-T sessions are the test data, where $T = 5, 6, 7, 8, 9, 10$. Here we randomly pick T sessions from the 12 sessions and consider all the possible combinations. The average prediction accuracy among all the combinations for each T is presented. Our algorithm is among those that achieve the highest predictive accuracy. Notice that while Randomized Ward Logistic reveals only a small number of discriminative voxels, its prediction accuracy is also very high. While high predictive accuracy does not directly prove the sensitivity or specificity of feature selection results, it still suggests the quality of the identification of voxels of different algorithm to some degree. At the least, the prediction

Algorithm \rightarrow	Ours	Rand Ward Logistic	TV-L1	Rand L1
Haxby	35	68	4	10
Chess-Master	52	433	15	36

Table 3: Running time (unit: minute) of different methods for second problem (Chess-Master Data) and the third problem (Haxby Cognitive Task Data, training sessions=10).

scores suggest that our algorithm does indeed find the relevant voxels because they can achieve significantly high predictive accuracy.

3.6. A Brief Computational Efficiency Description

All the above experiments were performed under Windows 7 and MATLAB_R2014a(V8.3.0.532) running on a desktop with Intel Core i7 Quad-Core (Eight-Thread) Processor with Processor Base Frequency 3.5GHz and 64 GB of memory, though there are no parallel implementations of all the involved algorithms. We listed the running time (unit of time is minute here) of different algorithms for the test problems based on both Chess-Master data and Haxby Cognitive Task Data in Table 3. Here we did not list the running of Smooth lasso and ℓ_2 -SVM, ℓ_2 logistic regression and T-test, because they in general take much shorter time than the listed 4 algorithms.

Notice that the random ward clustering algorithm is written in Python, while our algorithm is written in MATLAB. Python is in general a more computationally efficient computer language than MATLAB. So the advantage of our algorithm in terms of computational efficiency comparing with the random ward clustering algorithm is remarkable. Notice that for random ward clustering algorithm, its running time is significant longer for Chess-Master test problem than the Haxby test problem. The number of features of Chess-Master test problem is $91 \times 109 \times 91$ while the number of features of Haxby test problem is $40 \times 64 \times 64$. The spatially constrained ward clustering method used in the random ward clustering algorithm empirically takes a notably much longer time as the number of features increases. While both ours and random ward clustering algorithm take a longer time than the other alternative algorithms as expected, the running time is still acceptable in general. Finally, we would like to point out that for random ward clustering algorithm and TV-L1 algorithms, we directly use the default settings of the provided python software. Their computational efficiency could be much different if different parameters are used. So the presented running time of different algorithms here is only for a rough reference.

4. Conclusion and Future Work

Voxel selection is very important for decoding fMRI data. In this paper we propose a simple and computationally efficient method for data-driven voxel selection which is also called support identification, for potential biomarker extraction (Orrù et al., 2012). We propose a “*randomized structural sparsity*” as a structural variant of classical stability selection via specific implementation “*Constrained Block Subsamplings*”. We apply this to the existing sparse multi-variate classifiers such as ℓ_1 logistic regression, in the case of fMRI data, which has strong correlations and distributed multivariate discriminative patterns. However, the results are mostly empirical and we might need to perform theoretical support in order to better understand its advantages and address its limitations. For example, the theoretical results about the false positive rate and false negative rate of our feature selection algorithm need to be presented in the future work. In addition, we need to further study the possible bias or arbitraries brought by our one-time parcelation. Moreover, we would like to try the hierarchical ward clustering with spatially constraints in our algorithm in future. It has showed that in general Ward’s clustering performs better than alternative methods with regard to reproducibility and accuracy (Thirion et al., 2014). Furthermore, how to effectively distinguish true positives from false positives needs to be better addressed.

Acknowledgment

Thanks to Prof. Alexandre Gramfort of Telecom ParisTech for kindly providing us with the Python Smooth LASSO and TV-L1 code, which is under integration in the Nilearn package. Thanks to Dr. Gaël Varoquaux from Parietal team, INRIA, for kindly providing us with the Randomized Ward Logistic algorithm written in Python. We would also like to thank the anonymous reviewers for their many constructive suggestions, which have greatly improved this paper.

This work was supported by the 973 programs (Nos. 2015CB856000, 2012CB517901), 863 project (SQ2015AA0201497), the Natural Science Foundation of China (Nos. 11201054, 91330201, 61125304, 81301279), and the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20120185110028) and, the Fundamental Research Funds for the Central Universities (ZYGX2013Z004, ZYGX2013Z005)

References

- Aguirre, G., Zarahn, E., D'esposito, M., 1998. The variability of human, bold hemodynamic responses. *Neuroimage* 8, 360–369.
- Anderson, M.L., Oates, T., 2010. A critique of multi-voxel pattern analysis, in: *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012a. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.* 4, 1–106. doi:10.1561/22000000015.
- Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012b. Structured sparsity through convex optimization. *Statist. Sci.* 27, 450C468.
- Baldassarre, L., Mourao-Miranda, J., Pontil, M., 2012. Structured sparsity models for brain decoding from fmri data, in: *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pp. 5–8.
- Batmanghelich, N., Taskar, B., Davatzikos, C., 2012. Generative-discriminative basis learning for medical imaging. *Medical Imaging, IEEE Transactions on* 31, 51–69. doi:10.1109/TMI.2011.2162961.
- Beinrucker, A., Dogan, u., Blanchard, G., 2012. A simple extension of stability feature selection, in: Pinz, A., Pock, T., Bischof, H., Leberl, F. (Eds.), *Pattern Recognition. Springer Berlin Heidelberg*. volume 7476 of *Lecture Notes in Computer Science*, pp. 256–265.
- Beinrucker, A., Gogan, u., Blanchard, G., 2015. Extensions of stability selection using subsamples of observations and covariates. *Arxiv:1407.4916v2*.
- Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97, 245 – 271. URL: <http://www.sciencedirect.com/science/article/pii/S0004370297000635>, doi:[http://dx.doi.org/10.1016/S0004-3702\(97\)00063-5](http://dx.doi.org/10.1016/S0004-3702(97)00063-5). relevance.
- Breiman, L., 1996. Bagging predictors. *Machine Learning* 24, 123–140. doi:10.1023/A:1018054314350.

- Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32. doi:10.1023/A:1010933404324.
- Bühlmann, P., Rütimann, P., van de Geer, S., Zhang, C.H., 2013. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference* 143, 1835 – 1858. URL: <http://www.sciencedirect.com/science/article/pii/S0378375813001225>, doi:<http://dx.doi.org/10.1016/j.jspi.2013.05.019>.
- Bühlmann, P., Van De Geer, S., 2011. *Statistics for high-dimensional data: methods, theory and applications*. Springer.
- Cao, H., Duan, J., Lin, D., Shugart, Y.Y., Calhoun, V., Wang, Y.P., 2014. Sparse representation-based biomarker selection for schizophrenia with integrated analysis of fmri and {SNPs}. *NeuroImage*, –.
- Chen, X., Lin, Q., Kim, S., Carbonell, J.G., Xing, E.P., 2012. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics* 6, 719–752.
- Chi, Z., et al., 2008. False discovery rate control with multivariate p-values. *Electronic Journal of Statistics* 2, 368–411.
- Cover, T., 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers* 14, 326–334.
- Craddock, R., Jbabdi, S., Yan, C-G., e.a., 2013. Imaging human connectomes at the macroscale. *Nature methods* 10, 524–539.
- Dubois, M., Hadj Seleem, F., Lofstedt, T., Frouin, V., Duchesnay, E., 2014. Predictive support recovery with sparse-tv penalties and logistic regression: an application to structural mri URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6858517>.
- Duda, R.O., Hart, P.E., Stork, D.G., 2000. *Pattern Classification* (2Nd Edition). Wiley-Interscience.
- Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9, 1871–1874.

- Fellinghauer, B., Bühlmann, P., Ryffel, M., Von Rhein, M., Reinhardt, J.D., 2013. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis* 64, 132–152.
- Flandin, G., Kherif, F., Pennec, X., Malandain, G., Ayache, N., Poline, J.B., 2002. Improved detection sensitivity in functional mri data using a brain parcelling technique, in: Dohi, T., Kikinis, R. (Eds.), *Medical Image Computing and Computer-Assisted Intervention MICCAI 2002*. Springer Berlin Heidelberg. volume 2488 of *Lecture Notes in Computer Science*, pp. 467–474. URL: http://dx.doi.org/10.1007/3-540-45786-0_58, doi:10.1007/3-540-45786-0_58.
- Gaël Varoquaux, Alexandre Gramfort, B.T., 2012. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering, in: ICML.
- Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. *Neural Comput.* 4, 1–58. URL: <http://dx.doi.org/10.1162/neco.1992.4.1.1>, doi:10.1162/neco.1992.4.1.1.
- Gramfort, A., Thirion, B., Varoquaux, G., 2013. Identifying predictive regions from fmri with tv-l1 prior, in: *Proceedings of the 2013 International Workshop on Pattern Recognition in Neuroimaging*, IEEE Computer Society, Washington, DC, USA. pp. 17–20. URL: <http://dx.doi.org/10.1109/PRNI.2013.14>, doi:10.1109/PRNI.2013.14.
- Gramfort, A., Varoquaux, G., Thirion, B., 2012. Beyond brain reading: Randomized sparsity and clustering to simultaneously predict and identify, in: Langs, G., Rish, I., Grosse-Wentrup, M., Murphy, B. (Eds.), *Machine Learning and Interpretation in Neuroimaging*. Springer Berlin Heidelberg. volume 7263 of *Lecture Notes in Computer Science*, pp. 9–16.
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157–1182.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. Ensemble learning, in: *The Elements of Statistical Learning*. Springer New York. Springer Series in Statistics, pp. 605–624.

- Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430.
- Hebiri, M., van de Geer, S., 2011. The smooth-lasso and other $\ell_1 + \ell_2$ -penalized methods. *Electron. J. Statist.* 5, 1184–1226. URL: <http://dx.doi.org/10.1214/11-EJS638>, doi:10.1214/11-EJS638.
- Hofner, B., Boccuto, L., Göker, M., 2014. Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *arXiv preprint arXiv:1411.1285*.
- Huang, J., Zhang, T., 2010. The benefit of group sparsity. *The Annals of Statistics* 38, pp. 1978–2004. URL: <http://www.jstor.org/stable/20744481>.
- Jacob, L., Obozinski, G., Vert, J.P., 2009. Group lasso with overlap and graph lasso, in: *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, New York, NY, USA. pp. 433–440.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An Introduction to Statistical Learning with Applications in R*. volume 103 of *Springer Texts in Statistics*. Springer.
- Jenatton, R., Audibert, J.Y., Bach, F., 2011. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research* 12, 2777–2824.
- Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., Thirion, B., 2012. Multiscale mining of fmri data with hierarchical structured sparsity. *SIAM J. Imaging Sciences* , 835–856.
- Lahiri, S., 2001. Effects of block lengths on the validity of block resampling methods. *Probability Theory and Related Fields* 121, 73–97. URL: <http://dx.doi.org/10.1007/PL00008798>, doi:10.1007/PL00008798.
- Lahiri, S.N., 1999. Theoretical comparisons of block bootstrap methods. *Ann. Statist.* 27, 386–404.

- Langs, G., Menze, B.H., Lashkari, D., Golland, P., 2011. Detecting stable distributed patterns of brain activation using gini contrast. *NeuroImage* 56, 497 – 507. URL: <http://www.sciencedirect.com/science/article/pii/S1053811910010669>, doi:<http://dx.doi.org/10.1016/j.neuroimage.2010.07.074>. multi-variate Decoding and Brain Reading.
- Li, Y., Long, J., He, L., Lu, H., Gu, Z., et al, 2012. A sparse representation-based algorithm for pattern localization in brain imaging data analysis. *PLoS ONE* 7.
- Li, Z., Liu, J., Yang, Y., Zhou, X., Lu, H., 2013. Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering* 99, 1. doi:<http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.65>.
- Liu, J., Ji, S., Ye, J., 2009a. Multi-task feature learning via efficient $\ell_{2,1}$ -norm minimization, in: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, Arlington, Virginia, United States. pp. 339–348.
- Liu, J., Ji, S., Ye, J., 2009b. SLEP: Sparse Learning with Efficient Projections. Arizona State University. URL: <http://www.public.asu.edu/~jye02/Software/SLEP>.
- Liu, J., Ye, J., 2010. Moreau-yosida regularization for grouped tree structure learning, in: Lafferty, J., Williams, C., Shawe-taylor, J., Zemel, R., Culotta, A. (Eds.), *Advances in Neural Information Processing Systems* 23, pp. 1459–1467.
- Mairal, J., Yu, B., 2013a. Discussion about grouping strategies and thresholding for high dimensional linear models. *Journal of Statistical Planning and Inference* 143, 1451 – 1453. URL: <http://www.sciencedirect.com/science/article/pii/S0378375813000475>, doi:<http://dx.doi.org/10.1016/j.jspi.2013.03.002>.
- Mairal, J., Yu, B., 2013b. Supervised feature selection in graphs with path coding penalties and network flows. *J. Mach. Learn. Res.* 14, 2449–2485. URL: <http://dl.acm.org/citation.cfm?id=2567709.2567740>.

- Meinshausen, N., 2013. Discussion of grouping strategies and thresholding for high dimension linear models. *Journal of Statistical Planning and Inference* 143, 1439 – 1440. URL: <http://www.sciencedirect.com/science/article/pii/S0378375813000505>, doi:<http://dx.doi.org/10.1016/j.jspi.2013.03.005>.
- Meinshausen, N., Bühlmann, P., 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 417–473. URL: <http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x>, doi:10.1111/j.1467-9868.2010.00740.x.
- Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., Thirion, B., 2012. A supervised clustering approach for fmri-based inference of brain states. *Pattern Recognition* 45, 2041 – 2049.
- Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B., 2011. Total variation regularization for fmri-based prediction of behavior. *Medical Imaging, IEEE Transactions on* 30, 1328–1340. doi:10.1109/TMI.2011.2113378.
- Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to decode cognitive states from brain images. *Mach. Learn.* 57, 145–175.
- Mota, B.D., Fritsch, V., Varoquaux, G., Banaschewski, T., Barker, G.J., Bokde, A.L., Bromberg, U., Conrod, P., Gallinat, J., Garavan, H., Martinot, J.L., Nees, F., Paus, T., Pausova, Z., Rietschel, M., Smolka, M.N., Ströhle, A., Frouin, V., Poline, J.B., Thirion, B., 2014. Randomized parcellation based inference. *NeuroImage* 89, 203 – 215.
- Ng, B., Abugharbieh, R., 2011. Generalized group sparse classifiers with application in fmri brain decoding, in: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Washington, DC, USA. pp. 1065–1071.
- Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience & Biobehavioral Reviews* 36, 1140 – 1152. URL:

- <http://www.sciencedirect.com/science/article/pii/S0149763412000139>,
doi:<http://dx.doi.org/10.1016/j.neubiorev.2012.01.004>.
- Park, M.Y., Hastie, T., Tibshirani, R., 2007. Averaged gene expressions for regression. *Biostatistics* 8, 212–227. doi:10.1093/biostatistics/kx1002.
- Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? *Trends in cognitive sciences* 10, 59.
- Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C., 2012. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition* 45, 2085 – 2100.
- Rondina, J., Hahn, T., de Oliveira, L., Marquand, A., Dresler, T., Leitner, T., Fallgatter, A., Shawe-Taylor, J., Mourao-Miranda, J., 2014. Scors - a method based on stability for feature selection and mapping in neuroimaging. *Medical Imaging, IEEE Transactions on* 33, 85–98. doi:10.1109/TMI.2013.2281398.
- Ryali, S., Chen, T., Supekar, K., Menon, V., 2012a. Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage* 59, 3852–3861.
- Ryali, S., Chen, T., Supekar, K., Menon, V., 2012b. Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage* 59, 3852 – 3861. URL: <http://www.sciencedirect.com/science/article/pii/S105381191101336X>, doi:<http://dx.doi.org/10.1016/j.neuroimage.2011.11.054>.
- Särndal, C.E.e.a., 2003. *Model Assisted Survey Sampling*. Springer. chapter Stratified Sampling. pp. 100–109.
- Schmidt, M., Roux, N.L., Bach, F.R., 2011. Convergence rates of inexact proximal-gradient methods for convex optimization, in: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (Eds.), *Advances in Neural Information Processing Systems* 24. Curran Associates, Inc., pp. 1458–1466. URL: <http://papers.nips.cc/paper/4452-convergence-rates-of-inexact-proximal-gradient>.
- Shah, R.D., Samworth, R.J., 2013. Variable selection with error control: another look at stability selection. *J. R. Statist. Soc. B* 75, 55–80.

- Thirion, B., Varoquaux, G., Dohmatob, E., Poline, J.B., 2014. Which fmri clustering gives good brain parcellations? *Frontiers in Neuroscience* 8.
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in {SPM} using a macroscopic anatomical parcellation of the {MNI} {MRI} single-subject brain. *NeuroImage* 15, 273 – 289. URL: <http://www.sciencedirect.com/science/article/pii/S1053811901909784>, doi:<http://dx.doi.org/10.1006/nimg.2001.0978>.
- de Vries, P., 1986. Stratified random sampling, in: *Sampling Theory for Forest Inventory*. Springer Berlin Heidelberg, pp. 31–55.
- Witten, D.M., Shojaie, A., Zhang, F., 2014. The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics* 56, 112–122. doi:10.1080/00401706.2013.810174.
- Xia, Z., Zhou, X., Chen, W., Chang, C., 2010. A graph-based elastic net for variable selection and module identification for genomic data analysis, in: *Bioinformatics and Biomedicine (BIBM)*, 2010 IEEE International Conference on, pp. 357–362. doi:10.1109/BIBM.2010.5706591.
- Xiang, S., Shen, X., Ye, J., 2012. Efficient sparse group feature selection via nonconvex optimization. *CoRR* abs/1205.5075.
- Xiang, S., Shen, X., Ye, J., 2015. Efficient nonconvex sparse group feature selection via continuous and discrete optimization. *Artificial Intelligence* 224, 28 – 50. URL: <http://www.sciencedirect.com/science/article/pii/S0004370215000302>, doi:<http://dx.doi.org/10.1016/j.artint.2015.02.008>.
- Yamashita, O., Aki Sato, M., Yoshioka, T., Tong, F., Kamitani, Y., 2008. Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. *NeuroImage* 42, 1414–1429.
- Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V., for the Alzheimer’s Disease Neuroimaging Initiative, 2012. Sparse learning and stability selection for predicting mci to ad conversion using baseline adni data. *BMC Neurology* 12, 46.

- Yu, L., Ding, C., Loscalzo, S., 2008. Stable feature selection via dense feature groups, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 803–811. URL: <http://doi.acm.org/10.1145/1401890.1401986>, doi:10.1145/1401890.1401986.
- Yuan, L., Liu, J., Ye, J., 2013. Efficient methods for overlapping group lasso. Pattern Analysis and Machine Intelligence, IEEE Transactions on PP, 1–1.
- Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 49–67. URL: <http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x>, doi:10.1111/j.1467-9868.2005.00532.x.
- Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

This figure "abide_logo.jpg" is available in "jpg" format from:

<http://arxiv.org/ps/1410.4650v2>